Published in final edited form as:
J Proteome Res. 2012 August 3; 11(8): 4013–4023. doi:10.1021/pr300058z.

Mapping the Protein Domain Structures of the Respiratory Mucins: a mucin proteome coverage study

Rui Cao, **T. Tiffany Wang**, **Genevieve DeMaria**, **John K. Sheehan**, and **Mehmet Kesimer***
Cystic Fibrosis and Pulmonary Research Center, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, NC, USA

Abstract

Mucin genes encode a family of the largest expressed proteins in the human genome. The proteins are highly substituted with *O*-linked oligosaccharides which greatly restrict access to the peptide backbones. The genomic organization of the N-terminal, *O*-glycosylated, and C-terminal regions of most of the mucins has been established and is available in the sequence databases. However, much less is known about the fate of their exposed protein regions after translation and secretion, and, to date, detailed proteomic studies complementary to the genomic studies are rather limited. Using mucins isolated from cultured human airway epithelial cell secretions, trypsin digestion and mass spectrometry, we investigated the proteome coverage of the mucins responsible for the maintenance and protection of the airway epithelia. Excluding the heavily glycosylated mucin domains, up to 85% coverage of the N-terminal region of the gel forming mucins MUC5B and MUC5AC was achieved, and up to 60% of the C-terminal regions were covered, suggesting that more *N*- and sparsely *O*-glycosylated regions as well as possible other modifications are available at the C-terminus. All possible peptides from the cysteine-rich regions that interrupt the heavily glycosylated mucin domains were identified. Interestingly, 43 cleavage sites from ten different domains of MUC5B and MUC5AC were identified, which possessed a non-tryptic cleavage site on the *N*-terminal end of the peptide, indicating potential exposure to proteolytic and/or "spontaneous cleavages". Some of these non-tryptic cleavages may be important for proper maturation of the molecule, before and/or after secretion. Most of the peptides identified from MUC16 were from the SEA region. Surprisingly, three peptides were clearly identified from its heavily glycosylated regions. Up to 25% coverage of MUC4 was achieved covering seven different domains of the molecule. All peptides from the MUC1 cytoplasmic domain were detected along with the three non-tryptic cleavages in the region. Only one peptide was identified from MUC20 which led us to successful antisera raised against the molecule. Taken together, this report represents our current efforts to dissect the complexities of mucin macromolecules. Identification of regions accessible to proteolysis can help in the design of effective antibodies and points to regions that might be available for mucin-protein interactions and identification of cleavage sites will enable understanding of their pre- and post-secretory processing in normal and disease environments.

Keywords

Mucins; respiratory; proteomics; coverage; MUC5B; MUC5AC

*Correspondence Author: 4021 Thurston Bowles bldg. CB#7248, Chapel Hill NC, 27599, kesimer@med.unc.edu.

Introduction

Mucins are high molecular weight glycoproteins that are physically large in size, as well as high in mass. At least 20 mucin genes have been identified to date. At least ten human mucin genes (membrane tethered mucins MUC1, 4, and 16, and the gel forming mucins MUC5AC, 5B, as well as others not so well characterized in the airways e.g. MUCs 2, 7, 13, 15, 19 and 20) have been observed at the mRNA level in lung tissue from healthy individuals [1, 2]. In a previous report, we identified five of these mucin gene products using proteomic approaches on mucus harvested from primary human epithelial cell culture secretions, and induced sputum[3]. The five identified mucins were the gel forming mucins MUC5AC and MUC5B and the membrane mucins MUC1, MUC4 and MUC16. MUC20 was not reported because it is a relatively new member of the mucin family and its sequence was not available in the databases at the time of the publication.

The so called gel-forming mucins are essential for the formation of a flowing mucus gel vital for epithelial protection and lung function. In healthy lungs, MUC5B is mostly produced by mucous cells in the sub-mucosal glands[4] while MUC5AC is produced by airway surface goblet cells [5]. In disease [6] and airway epithelial cell cultures [7], however, MUC5B is produced by surface goblet cells as well. The other three well-characterized mucins all belong to the transmembrane family; MUC1 is most strongly identified with the microvilli, while MUC4 appears to be strongly associated with the cilia [8]. MUC16 is more mysterious, as it is strongly associated with MUC5B secreting cells in the sub-mucosal glands; however, in cell culture, it is expressed in goblet cells [8], which secrete MUC5B, predominantly. Both gel forming and membrane related mucins of the respiratory tract function as part of the airway's innate defense system by enabling the capture and elimination of particulates and pathogens by mucociliary transport and/or cough clearance.

As a protein product of MUC genes, mucins show complex multi-domain structures. For example, the large gel forming mucins have at least two serine, threonine, proline (STP) rich, heavily glycosylated mucin domains (MD) with small cysteine-rich compartments localized within these domains as well as two to four Von Willebrand Factor (vWF) like domains situated at the N- and C- terminal regions [9,10]. Membrane related mucins of the airways also show complexity by displaying multi domain structures such as VNTR (variable number of tandem repeats), *O*-glycan rich domains, *N*-glycosylated regions, cysteine-rich regions, SEA regions, and transmembrane, and cytoplasmic domains [1, 11, 12]. Membrane tethered mucins can also be found as secreted soluble forms released from the cell surface after protein cleavage or, alternatively, as spliced secreted forms sometimes lacking their distinctive glycosylated domains. The identification of the gene sequences of these molecules has now been achieved and the genomic organization of N-terminal, *O*-glycosylated mucin domains and C-terminal regions of MUC5B and MUC5AC has been well characterized [9, 10] and is available in sequence databases. However, much less is known about the state form and function of the naked protein regions themselves, and proteomic studies complementary to the genomic information are rather limited. Information such as the naked protein regions accessible and/or inaccessible to proteolytic cleavages, cleavage products, and post translational modifications are all crucial to understanding the three dimensional protein structure and structure/function relationships. This information is also important for the practical purpose of producing effective antibodies, to reveal interaction sites and understanding more about their biological roles in health and in the progression of chronic airway diseases. For these reasons, we have undertaken a major proteomic study of airway mucins and now report the most detailed peptide coverage available, which promises to advance our understanding of these large and complex macromolecules.

Experimental

Cell culture

Two different airway cell culture systems that secrete mucus were used in this study as a source of mucins: human tracheobronchial epithelial cells (HTBE) and airway epithelial Calu-3 cells. HTBE cells in cultures produces more MUC5B than MUC5AC (up to 10-fold) [13]. Additionally, HTBE cells are a good source for membrane tethered mucins [3]. Previous studies suggest that Calu-3 cells are enriched for goblet (mucous) cells [14] and our unpublished studies indicate that these cells produce more MUC5AC than MUC5B (up to 5 fold). For this reason, HTBE secretions were used as a source of MUC5B as well as all membrane tethered mucins and Calu-3 secretions were used for MUC5AC coverage in this report. HTBE cells were obtained from airway specimens resected from normal donor tissue from the University of North Carolina (UNC) lung transplant program or from NDRI (National Disease Research Interchange) under UNC Institutional Review Board-approved protocols. Primary airway epithelial cells were isolated by the UNC Cystic Fibrosis (CF) Center Tissue Culture Core and expanded on plastic to generate passage 1 cells and plated at a density of 600 k cells per well on permeable Transwell-Col (T-Col, 24 mm diameter) supports. Human tracheobronchial epithelia (HTBE) cultures were generated by provision of an air-liquid interface for 4–6 weeks to form well-differentiated, polarized cultures that resemble *in vivo* pseudo-stratified mucociliary epithelium[15]. Airway epithelial Calu-3 cells were derived from pleural effusion associated with a human lung adenocarcinoma. Calu-3 cells were grown on 12-mm Transwell supports (Corning) and maintained at air–liquid interface for at least three weeks, as previously described[14]. Mucus secretions from both cell cultures were obtained by incubating 1 ml of PBS on the apical surface of HTBE and Calu-3 cultures for 30 min at 37°C, removing the PBS with a large caliber pipette and then repeating the procedure one time. The 2 ml of diluted mucus secretions obtained per culture were pooled and placed immediately on ice until solubilization with solid GuHCl to make 4M solution..

Sample preparation

Mucin samples were prepared as previously described[3]. Briefly, collected secretion material was originally in PBS. An appropriate amount of solid GuHCl was then added to that material to make a 4M GuHCl solution, then CsCl was added to a density of 1.45 g/ml. Isopycnic density-gradient centrifugation was performed for 60 h at 36000 rpm on a Beckman L8-M ultracentrifuge using 50.2TI, 12 × 40 ml rotor. The sample was unloaded as 2 ml fractions from the top and then analyzed for PAS staining, antibody, and density. PAS-rich fractions (8–13) were pooled as the mucin rich fraction. High density regions (fractions 14–20) were pooled and analyzed separately as MUC4 and MUC16 rich regions. The pools were reduced by adding 10 mM dithiothreitol (DTT) to each sample for 2–3 h at 37°C. Free thiols were subsequently alkylated by the addition of 30 mM iodoacetamide for 1 h in the dark at ambient temperature. 1.5 ml of the reduced and alkylated samples was subjected to a HiTrap Desalting column (Sephadex G-25, Amersham Biosciences) to exchange the buffer to 50 mM ammonium hydrogen carbonate (NH_4HCO_3). Two ml of eluents were collected. 0.5 μg modified trypsin (proteomics grade, Sigma) was added to samples then incubated 18 h at 37°C. Three independent digests of the same pool were prepared. Tryptic digests (peptides and glycopeptides) were chromatographed on an Amersham Biosciences Superdex 200 HR 10/30 column using Ettan *LC* (Amersham Pharmacia Biotech) chromatographic equipment. The column was eluted with 50 mM NH_4HCO_3 at a flow rate of 0.3 ml/min. The void (large glycopeptides) and the including (peptides) volume were collected separately. The peptide pool was dried down ten times by volume using a Heto vacuum concentrator and samples were mixed 1:1 with 1% formic acid and then subjected to nano-LC-MS/MS analysis. To determine N- glycosylation sites, an aliquot from the reduced and alkylated

pools was incubated with PNGase F (Sigma) for 60 minutes at 37°C and subjected to digestion as described above (digest #3).

Mass Spectrometry/Electrospray LC-MS/MS

Digested samples (2μL) were introduced via a Waters nanoACQUITY UPLC system. The analytical system was attached directly to the Z spray source and was configured with a PepMap™ C18 (LC Packing, 300Nm ID × 5mm) pre-concentration column in series with an Atlantis® (Waters) dC18 NanoEase™ (75 m x 150 mm) nanoscale analytical column. The spray tip used was PicoTip™ (New objectives USA) capillary with 10 μm diameter. The pump was programmed to deliver 0.4 μL/min, through the analytical column. Samples were separated on the column with a gradient of 5% acetonitrile in 0.1% formic acid to 60 % acetonitrile in 0.1% formic acid over 120 min. The composition of the solvents A, B and C was 5% (v/v) ACN in 0.1% (v/v) formic acid, 95% (v/v) ACN in 0.1% (v/v) formic acid and 0.1% (v/v) formic acid, respectively. All data were acquired using Waters Q-Tof *micro*, hybrid quadropole orthogonal acceleration time-flight mass spectrometer (Waters, Manchester, UK) with MassLynx 4.1 software. The mass spectrometer data directed analysis (DDA) acquired MS survey data from m/z 200 to 1500 with the criteria for MS to MS/MS including ion intensity and charge state using a one second MS survey scan followed by 1.5 second MS/MS scans, each on three different precursor ions. The Q-Tof micro was programmed to ignore any singly charged species and the collision energy used to perform MS/MS was carried out according to the mass and charge state of the eluting peptide. Precursors detected were excluded from any further MS/MS experiment for 180 seconds. All analyses were repeated three times for each sample, and peptides identified in the first run were excluded from the second analysis.

Data processing and database searching

The raw data acquired were processed using the ProteinLynx module of MassLynx 4.0 to produce *.pkl (peaklist) files which are suitable for the MS/MS ions database search via search engines. The peptide QA filter was 30 to eliminate poor quality spectra and the minimum peak width at half height was set to four to eliminate background noise peaks. Smoothing (x2 Savitzky Golay) and polynomial fitting were performed on all peaks and the centroid taken at 80% of the peak height. The processed data was searched against the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein database (version 2011-09-12: 15,270,974 sequences, 4,841,198 Homo sapiens sequences) and Swiss-Prot (version 2011-09-21: 532,146 sequences, 20,249 Homo sapiens sequences) using the in-house MASCOT (Matrix Science, UK) search engine (Version 2.0). Parameters used for the MASCOT search were: Taxonomy human, 0.2 Da mass accuracy for parent ions and 0.3 Da mass accuracy for fragment ions, one missed cleavage was allowed, and carbomidomethyl-Cys and methionin oxidation were used as fixed and variable modifications, respectively. MASCOT probability-based Mowse individual ion scores > 40 were accepted as indicating identity or extensive homology ($p<0.05$). MS/MS spectrum scores between 20–40 were examined individually with the acceptance criteria that the parent and fragment ion masses were within the calibrated tolerance limits and that the spectrum contained an extended series of consecutive y- or b- ions.

Detecting half/non-tryptic peptides

A Mascot “error tolerant” search was performed to examine unmatched spectra resulting from non-tryptic/non-specific cleavages, mass measurement error, incorrect determination of precursor charge, unsuspected chemical & post-translational modifications, and other factors. No cleavage enzyme is specified in this stage of search, thereby allowing peptides originating from unexpected enzyme activity or self cleavage of the molecule to be identified.

Results and Discussion

Peptide coverage analysis of MUC5B

The MUC5B apoprotein (Swiss-Prot Accession # Q9HC84) has a large monoisotopic mass of 588,003 Da and yields 293 theoretical tryptic cleavage sites (ExPasy-PeptideCutter). The MD region has 79 tryptic peptides, all of which are heavily glycosylated. No peptides were identified from the MD region and the size of the glycopeptides recovered (Mw 0.1 MDa to 1 MDa) suggests that none of the potentially cleavable sites in the region could be accessed. Excluding the large central mucin domains where extensive glycosylation limits access by trypsin, about 181 of these are detectable by mass spectrometry of which 109 were successfully identified from 17 different regions of the protein (table 1, supplementary data 1). From the N-terminus, 52 of the 67 predicted tryptic peptides (85% coverage) were identified. All of the 35 predicted peptides from seven cysteine-rich regions inside the MD region were successfully identified. Of 34 peptides identified in the C-terminus, 22 (60% coverage) were identified. Some 35 peptides are composed of one to five amino acid residues, while eight peptides outside the MD region are larger than 5000 Da in mass, with five of these eight peptides having potential *O*-glycosylation sites (NetOGlyc 3.1: http://www.cbs.dtu.dk/services/NetOGlyc). The range of our acquisition was between 150–2000 m/z during a survey scan and 100–1500 m/z in the MS/MS mode; therefore, this acquisition was appropriate for the detection of peptides of up to 6000 m/z in mass employing triply charged peptides and 4000 for the doubly charged. Peptides that were smaller than approximately five amino acid residues were not within our accurate range of peptide identification.

The mature, fully glycosylated, MUC5B protein is relatively low in putative *N*-glycosylation sites and most of these are outside the MD region. Whether they are occupied is not clear. The C-terminus of the molecule has 12 putative *N*-glycosylation sites (NetNGlyc 1.0 Server: http://www.cbs.dtu.dk/services/NetNGlyc/), while only four putative *N*-glycosylation sites are available at the N-terminus. Only one extra peptide from the N-terminal region, ^{200}YA**NQT**CGLCGDFNGLPAFNEFYAHNAR^{225}, and two peptides from the C-terminal region, ^{5208}LPYSLFH**NNT**EGQCGTCTNNQR^{5229} and ^{5036}VVLLDPKPVA**NVT**CVNK^{5552} were identified after PNGase-F treatment which cleaves the *N*-linked glycans on the protein, which were consistent with this particular mode of glycosylation. Absence of other putative *N*-glycosylated peptides by PNGase F suggests that those peptides could be modified in some other way. Another putative site of glycosylation on mucins is mannosylation of the first tryptophan residue of the WXXW motif [16], called *C*-mannosylation. Each of the cysteine-rich domains contains one WXXW motif; however, all three predicted tryptophan mannosylation sites in the first three cysteine-rich domains were identified without any deglycosylation pre-processing. These peptides were ^{1340}**WSSW**YNGHR^{1348}, ^{1507}CQ**WTEW**FDEDYPK^{1520}, and ^{1787}CEWTEWFDV…ENIR^{1534}. Our observations indicate that those peptides were available for mass spectrometry identification without any modification, though it is not possible to rule out the presence of this modification at a low level. The other four predicted *C*-mannosylation sites are on peptides partly inside the MD/VNTR region but stretching some way into cysteine-rich domains four-seven. The N-terminal part of these peptides could be heavily glycosylated and thus, their identification would not be possible with our standard methods.

Cysteine subdomains (Cys1-Cys7) that interrupt the highly *O*-glycosylated MD regions consist of 108 amino acid residues, each containing ten cysteine residues in which intramolecular disulfide bonds can be formed. The functional significance of these regions in the middle of heavily glycosylated regions is still speculative, but a recent suggestion is that the CysD domains of MUC2 mucin might serve as non-covalent cross-links in the colonic mucus gel [17]. Large portions of the heavily glycosylated regions are present both

upstream and downstream of these cysteine-rich regions. We identified almost all technically suitable peptides from all cysteine-rich regions indicating the reduction process was suitable for breaking all of the disulfide bonds in these regions, thus making them accessible to proteases. Taken together, we have 30% coverage of the mature glycosylated MUC5B. However, after subtracting non-accessible heavily glycosylated MD regions, the *O*- and/or *N*-glycosylated regions and the tryptic fragments unidentifiable for technical reasons of the mass spectrometry, we cover the molecule up to 95 percent. Only five percent of the other readily identifiable peptides were missed from our exhaustive analysis, the majority of which were from the N-terminus. The inability to achieve 100 percent coverage suggests that the access of trypsin is restricted in some portions of the N- and C- terminal regions by other post-translational modifications. A previous study reported peptide coverage of MUC5B isolated from saliva. Using mass spectrometry, they identified 84 peptides from different domains of MUC5B [18].

Unexpectedly, we found 27 non-tryptic cleavage sites from MUC5B, 16 of which were from the N-terminal region (table 4). In particular, four cleavage site were from inside the vWFD domains, ten from the unique region, one from the TIL region, and one from the MUC11p15-type region. ^{204}CGLCG^{208} and ^{1023}CGLCG^{1027} repeat sequence in both vWD1 and vWD3 regions are highly conserved in secreted mucins and have been proposed to be essential for mutimerization at low pH in the trans-Golgi [19]. Interestingly, one of the cleavage sites we identified was at the D1 domain and divides the CGLCG disulfide isomerase catalytic site (YA**NQT**CGL//^{1026}CGDFNGL……YAHNAR^{1044}). More interestingly, the sequence immediately upstream of the cleavage has a putative *N*-glycosylation (…NQT…) site; this cleavage site could be detected only after PNGase F treatment, suggesting possible protection of this multimerization site by *N*-glycosylation. We also found three different cleavage sites inside the Cys-rich regions: the VL// CCSDDHCR peptide found inside the second cysteine-rich region, ELGQVVEC// SLDFGLVCR which was present in the four through seven cysteine-rich regions as a four times repeat peptide and ^{4176}AQAQ//PGVPLGELGQVVECSLDFGLVCR^{4202}, which was found in the last cysteine-rich domain. The other five cleavage sites were from the C-terminal end (i.e. one from MUC11p15 region, one from unique region, one from vWFD domain and two from vWFC domain) (table 4). The MS/MS spectra of all non-tryptic cleavage products were supplied as supplementary data (supplementary data 2).

Two distinct proteolytic cleavage regions on salivary MUC5B—one cleavage in the D′ region and one in the D3 domain—and their importance have been previously proposed [20]. The cleavage described in D′ region is similar to those found in vWF protein [21] that cleaves a fragment of the N-terminal region up to D′ domain and may represent an important step for post-secretory processing. However, our coverage analysis on MUC5B indicates that the entire N-terminal region, including the D1, D2, D′, and D3 domains can be isolated and identified, to be part of a mature and fully glycosylated mucin suggesting the molecule is physically intact. The other cleavage, in D3, cleavage determines structural differences between soluble and insoluble MUC5B in saliva[22]. The non-tryptic cleavage sites we identified in this portion of the N-terminal region might be putative sites for such cleavages but this remains to be validated in future studies.

Peptide coverage analysis of MUC5AC

The MUC5AC apoprotein (Swiss-Prot # P98088) is about 496734 Da in Mw (the complete genomic sequence for the central MD regions is not completely clear) and yields 197 tryptic fragments after digestion. Of these, 138 peptides were identified from 22 different regions (Table 2, supplementary data 1). The longest peptide identified was 48 amino acids long with a Mw of 5008.54 Da. Of these detected tryptic peptides, 54 (77% coverage) were from

the N terminus. All 64 (100%) tryptic peptides of at least six residues or longer were identified from nine different cysteine-rich domains within the central MD regions. Out of a total of 35 tryptic peptides in the C-terminus region, 20 (60% coverage) were identified. The MD region, as so far described, consists of only 18 peptides, all of which are heavily glycosylated. From a mass spectrometry recovery view, two peptides are larger than 5000 m/z and 76 fragments are between one and five residues long. There are 20 predicted tryptic STP-rich peptides between the first five Cys-rich domains. These regions are found upstream of the large MD region and are heavily glycosylated even though they have no repeat sequences. There are three putative *N*-glycosylation sites at the N-terminus and seven at the C-terminus. Two *N*-glycosylation sites after PNGase F deglycosylation were found, both of which were from the C-terminal region. The peptide ^{4431}FA**NNT**EGQCGTCTNDR^{4446} is from the vWFD4 domain while the other, ^{4963}TSLR**NVT**LHCTDGSSR^{4978} is located in the cysteine-knot region. As with MUC5B, all possible predicted peptides inside the small cysteine-rich regions inside the heavily glycosylated MD region were identified. Like MUC5B, all of the nine cysteine-rich regions have one WXXW motif. The peptide CT**WTTW**FDVDFPSPGPHGGDK inside the fifth Cyst region was identified and the five times repeated motif WTKW has the trypsin cleavage site (WTK/W) of which the C-teminal part of the peptide was identified, but the N-terminal side was unidentified. This is possibly due to the *C*-mannosylation and/or *O*-glycosylation of the region. We have less coverage of the C-terminus of MUC5AC as compared with MUC5B. This is perhaps due to more *N*- and/or sparsely *O*-glycosylated sites or to other post-translational modifications available at that region. Overall, we have 40 percent coverage of the whole MUC5AC protein which can be up to 90 percent after excluding the heavily glycosylated MD region, the predicted *N*- and sparsely *O*-glycosylated peptides and the peptides beyond our detection limits. An ‘error tolerant’ analysis of MUC5AC yielded 16 non-specific cleavage peptides. Of these, seven were from N-terminal unique regions and were two from the TIL domain. Unlike MUC5B, no non-tryptic cleavage was found inside D domains at the N-terminal region (table 4, supplementary data 2). We also identified five different non-specific cleavage sites on two predicted tryptic peptides, one from the C-terminal vWFD domain and the other from the N-glycosylation rich region. Although a cleavage in the ^{4301}**GDPH**^{4304} sequence of the C-terminal region of MUC5AC mucin has been reported [23], we did not detect this particular cleavage.

Overall, the reason for the presence of an unexpectedly large quantity of non-tryptic cleavage sites on gel forming mucins MUC5B and MUC5AC and the complexity behind it is another mystery. Some of the non-tryptic cleavage sites are similar to chymotrypsin, NTCB, LysC, ArgC, and Asp-N type cleavages; however, most of these cleavage sites are non-specific in respect to any known cutting agents. As can be envisioned from figure 1, not every proteolytic cleavage can break mucin into smaller pieces since they are essentially held together by disulfide bonds throughout the molecule. Despite these cleavage sites on their naked protein domains, it appears that the gel forming mucins maintain their integrity and molecular composition as assessed by light scattering measurements (not shown) and agarose gel western blotting (supplementary figure 1). In abnormal conditions such as infection and inflammation, however, the number of cleavage sites may increase due to increased protease activity, which is either coming from host or intruder cells, and the mucin molecules can lose their integrity. Taken together, cleavages reported here could be a useful reference data for studies that use in-vivo secretions. For instance, comparison of the cleavages of mucins isolated from different sources (e.g. MUC5B from sputum, saliva, and cervical mucus under normal and disease conditions) will help better understanding of the functional significance of these cleavages and of the functional requirements of different mucosal surfaces. This should be to be pursued and elucidated in future studies.

Peptide coverage analysis of Membrane tethered mucins of airways

MUC20—MUC20 (Q8N307) is one of the smallest airway mucins. The expressed protein is about 72 kDa in mass. It has about 42 tryptic cleavage sites available in total and from these about three peptides may be identifiable in the protein sequences of the C-terminus region. Only one peptide ^{667}**LSVASPEDLTDPR**^{679} was detected and it is from the far C-terminal region close to the transmembrane domain.

MUC4—The molecular weight and size of the MUC4 (Q99102) molecule largely depends on the number of repeats of three tandem repeat regions in the MD regions of the N-terminal part. Taking one repeat, the MUC4 core protein has a predicted monoisotopic mass of 231,305 Da; however, given the range of 145–395 repeats in these regions, the actual varies from 5 to 15 MDa. Of a potential 78 available tryptic peptides outside the heavily glycosylated VNTR region, 32 peptides were identified from seven different domains of the C terminus region. Of these, 14 are from the unique sequence (CT2 and CT8) regions, ten from the cysteine-rich (CT3 and CT6) region, three from the vWD domain region (CT4), two from the *N*-glycosylation rich region (CT5) and three from the EGF2 domain (CT-9) (table 3, supplementary data 1). No extra peptide was found after PNGaseF treatment even though MUC4 has more than ten putative *N*-glycosylation sites mostly in the *N*-glycosylation region (CT5). Up to 35 percent coverage of the C terminus region was achieved and after subtracting the predicted *N*-glycosylation and technically unavailable peptides, that coverage ratio increased to up to 62 percent. Like MUC5AC, MUC4 also has a ^{1443}GDPH^{1446} cleavage site inside the vWFD domain though we could not identify the cleavage product. On the other hand, two non-specific non-tryptic cleavage sites were detected from the vWFD region, ^{1501}SSSL//^{1505}GPVTVQWLLEPHDAIR and ^{1501}SSSLG//^{1506}PVTVQWLLEPHDAIR (table 4). The MUC4 gene can generate up to 24 splice variants by alternative use of exons. Consequently, a question arises regarding the splice variants we have in our samples. According to our gel chromatographic and/or agarose gel electrophoresis experiments, we have identified only the mature glycosylated molecule in our cell culture mucus samples (supplementary figure 1). Evidence from the coverage mapping from five different CT regions suggests this material is the sv0 form of MUC4 with no splicing[24].

MUC1 and MUC16—The C-terminal extracellular regions of MUC1 and MUC16 are relatively simple in comparison to those of MUC4. The extracellular region of MUC1 protein consists of an *O*-glycosylation rich VNTR sequence and a SEA (**S**ea urchin sperm protein, **E**nterokinase and **A**grin) domain region. The membrane region of the protein contains transmembrane and cytoplasmic domains. The MUC1 protein (P15941) has a monoisotopic mass of 122,220 and yields only 19 tryptic peptides including one large VNTR residue (929 amino acids long and serine and threonine-rich) with no cleavage site. This region is followed by three sparsely glycosylated peptides, one of which exceeds 6000 m/z in mass, after which there is a SEA domain. There are only five identifiable peptides in this region of which we identified three; the other two, ^{1090}FRPGSVVVQLTLAFR^{1108} and ^{1156}TEAASR^{1161}, were not detected (table 3, supplementary data 1). The first of these sites contains a known FRPG//SVVV cleavage site [25] which may explain its absence. A large predicted *N*-glycosylated peptide (5961 m/z) follows the sequence, and the entire transmembrane part of the molecule is present. The last part of the molecule is the cytoplasmic tail and contains five tryptic peptides. Interestingly, we identified all five peptides covering 100% of this cytoplasmic region which is consistent with the absence of post-translational modifications. As we have previously described, we know that MUC1 is found on some populations of exosome-like vesicles[26]. Therefore, it is not surprising that we find its cytoplasmic domain. We also have evidence of non-tryptic cleavage sites inside the cytoplasmic domain, DTYHPMS//^{1208}EYPTYHTHGR^{1217}, MSEY//^{1210}PTYHTHGR and

YV//[1220]PPSSTDR[1226] (table 4). The cytoplasmic domain of MUC1 is a target for several kinases [27] and is involved in a number of active processes in the cell. The three non-specific cleavages are next to two putative phosphorylation sites by ZAP-70 and PKC gamma [27]. The functional significance of these cleavages inside the cytoplasmic domain is not clear; however, given the role of CT domain in transcriptional activity of the molecule via nuclear beta-catenin[28], one speculation is that these cleavages in the CT domain may regulate the transcription of the molecule.

MUC16 (Q8WXI7) is one of the largest expressed proteins in the genome with an expressed apoprotein has over 22,000 amino acids and has a mass of around 2.3 Mda. The extracellular repeat domain of the MUC16 molecule contains up to 60 tandem repeats of 156 amino acid long repeats [29]. A large proportion of the apoprotein, the first 12,000 amino acids, is the threonine and serine rich region and is followed by a large SEA region that contains over 120 small SEA domains (about 120 amino acids each) (Table 3, supplementary data 1). Over 100 of the peptides we found—virtually all of the peptides we identified—were from this repeat region. In this region, some 15 peptides repeat two to ten times, while 32 other peptides cover 30 percent of the repeat region. However, unlike any of the other mucins studied, three peptides, [5131]EIFPSINTEETNVK[5144], [6402]AFTAATTEVSR[6412] and [7578]VSTGATTEVSR[7588], were clearly identified within the serine/threonine rich glycosylated domain (Table 3, supplementary data 1). We have 50 percent coverage of the repeat region. Though MUC16 has about 300 putative *N*-glycosylation sites, we found only two *N*-glycosylated repetitive peptides after PNGase F treatment, LYWELS**NLT**NGIQELGPYTLDR and **NTS**VGLLYSGCR, suggesting most of the N-glycosylated regions were also O-glycosylated as well. The last portion of its C-terminal cytoplasmic domain, [22152]EGEYNVQQQCPGYYQSHLDLEDLQ[22175] was also clearly detected. Two repetitive non-tryptic cleavages were detected inside the SEA region SEKDGA//ATGVDAICTHR and LYWE//LSQLTHGIKELGPYTLDR. The GSVVV motif inside the SEA domains of membrane mucins, such as MUC1, MUC3 and MUC12, is the site for the cleavage of these molecules to release the N-terminal region to the mucus layer above [3031]. However, MUC16 has no such motif over its SEA region and no cleavage sites for MUC16 have been identified. The non-tryptic cleavages over the SEA region we presented here may be the potential cleavage sites for such a function.

In conclusion, we report here the most exhaustive proteome analysis of the major mucins released in the airways. The mucins described here are complex in both chemical structure and architecture. Not only do they have extended carbohydrate rich domains but also large unglycosylated and/or sparsely glycosylated protein rich regions that have definable tertiary structures. It is quite possible that much of this complexity is associated with specific biological functions such as binding other proteins, bacteria, viruses or other particulates or with locating specific protective proteins in the surface epithelium, but all of this is as yet undescribed in detail. This peptide coverage mapping of different domains and regions of the mucins gives us useful general information on their peptide accessibility, cleavage sites, post-translational modifications, and structure/function relationships. An important value of this information to us has been to understand the post-secretory destiny of gel forming mucins of the airways, MUC5B and MUC5AC, and to raise effective peptide antibodies for fully glycosylated transmembrane mucins, e.g. MUC4 and MUC20, whereas previously, only C-terminal cleavage sequences or non-glycosylated immature intracellular apoprotein could be experimentally identified. Unidentified peptides/regions outside the MD regions and the putative cleavage sites form an interest for our future studies related to post-secretory mucin maturation/processing and to identification and characterization of the post-translational modifications, and interaction sites for binding to one other and to other proteins available in the same environment. Understanding the complex multi domain structure of these huge gene products will also help to elucidate how these giant molecules

contribute to the airway's innate defense and to the progression of chronic airway diseases and will eventually help in designing better therapeutic modalities targeting the airway mucus.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Prof. C. William Davis for critical reading of the manuscript, the UNC CF Center Tissue Procurement and Culture Core for supplying the cell cultures, and Drs. Jack D. Griffith and Sezgin Ozgur for their help for electron microscopy. This work was supported by National Heart, Lung, and Blood Institute/National Institutes of Health (NHLBI/NIH) grant R01HL103940 (MK).

References

1. Hattrup CL, Gendler SJ. Structure and function of the cell surface (tethered) mucins. Annu Rev Physiol. 2008; 70:431–57. [PubMed: 17850209]
2. Thornton DJ, Rousseau K, McGuckin MA. Structure and function of the polymeric mucins in airways mucus. Annu Rev Physiol. 2008; 70:459–86. [PubMed: 17850213]
3. Kesimer M, Kirkham S, Pickles RJ, Henderson AG, Alexis NE, Demaria G, Knight D, Thornton DJ, Sheehan JK. Tracheobronchial air-liquid interface cell culture: a model for innate mucosal defense of the upper airways? Am J Physiol Lung Cell Mol Physiol. 2009; 296 (1):L92–L100. [PubMed: 18931053]
4. Sharma P, Dudus L, Nielsen PA, Clausen H, Yankaskas JR, Hollingsworth MA, Engelhardt JF. MUC5B and MUC7 are differentially expressed in mucous and serous cells of submucosal glands in human bronchial airways. Am J Respir Cell Mol Biol. 1998; 19 (1):30–7. [PubMed: 9651178]
5. Reid CJ, Gould S, Harris A. Developmental expression of mucin genes in the human respiratory tract. Am J Respir Cell Mol Biol. 1997; 17 (5):592–8. [PubMed: 9374110]
6. Groneberg DA, Eynott PR, Oates T, Lim S, Wu R, Carlstedt I, Nicholson AG, Chung KF. Expression of MUC5AC and MUC5B mucins in normal and cystic fibrosis lung. Respir Med. 2002; 96 (2):81–6. [PubMed: 11860173]
7. Bernacki SH, Nelson AL, Abdullah L, Sheehan JK, Harris A, Davis CW, Randell SH. Mucin gene expression during differentiation of human airway epithelia in vitro. Muc4 and muc5b are strongly induced. Am J Respir Cell Mol Biol. 1999; 20 (4):595–604. [PubMed: 10100990]
8. Sheehan JK, Kesimer M, Pickles R. Innate immunity and mucus structure and function. Novartis Found Symp. 2006; 279:155–66. discussion 167–9, 216–9. [PubMed: 17278393]
9. Buisine MP, Desseyn JL, Porchet N, Degand P, Laine A, Aubert JP. Genomic organization of the 3′-region of the human MUC5AC mucin gene: additional evidence for a common ancestral gene for the 11p15. 5 mucin gene family. Biochem J. 1998; 332 (Pt 3):729–38. [PubMed: 9620876]
10. Desseyn JL, Buisine MP, Porchet N, Aubert JP, Laine A. Genomic organization of the human mucin gene MUC5B. cDNA and genomic sequences upstream of the large central exon. J Biol Chem. 1998; 273 (46):30157–64. [PubMed: 9804771]
11. Lan MS, Batra SK, Qi WN, Metzgar RS, Hollingsworth MA. Cloning and sequencing of a human pancreatic tumor mucin cDNA. J Biol Chem. 1990; 265 (25):15294–9. [PubMed: 2394722]
12. Moniaux N, Nollet S, Porchet N, Degand P, Laine A, Aubert JP. Complete sequence of the human mucin MUC4: a putative cell membrane-associated mucin. Biochem J. 1999; 338 (Pt 2):325–33. [PubMed: 10024507]
13. Holmen JM, Karlsson NG, Abdullah LH, Randell SH, Sheehan JK, Hansson GC, Davis CW. Mucins and their O-Glycans from human bronchial epithelial cell cultures. Am J Physiol Lung Cell Mol Physiol. 2004; 287 (4):L824–34. [PubMed: 15194565]
14. Kreda SM, Okada SF, van Heusden CA, O'Neal W, Gabriel S, Abdullah L, Davis CW, Boucher RC, Lazarowski ER. Coordinated release of nucleotides and mucin from human airway epithelial Calu-3 cells. J Physiol. 2007; 584 (Pt 1):245–59. [PubMed: 17656429]

15. Fulcher ML, Gabriel S, Burns KA, Yankaskas JR, Randell SH. Well-differentiated human airway epithelial cell cultures. Methods Mol Med. 2005; 107:183–206. [PubMed: 15492373]
16. Kirkham S, Kolsum U, Rousseau K, Singh D, Vestbo J, Thornton DJ. MUC5B is the major mucin in the gel phase of sputum in chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2008; 178 (10):1033–9. [PubMed: 18776153]
17. Ambort D, van der Post S, Johansson ME, Mackenzie J, Thomsson E, Krengel U, Hansson GC. Function of the CysD domain of the gel-forming MUC2 mucin. Biochem J. 2011; 436 (1):61–70. [PubMed: 21338337]
18. Rousseau K, Kirkham S, Johnson L, Fitzpatrick B, Howard M, Adams EJ, Rogers DF, Knight D, Clegg P, Thornton DJ. Proteomic analysis of polymeric salivary mucins: no evidence for MUC19 in human saliva. Biochem J. 2008; 413 (3):545–52. [PubMed: 18426393]
19. Perez-Vilar J, Hill RL. Identification of the half-cystine residues in porcine submaxillary mucin critical for multimerization through the D-domains. Roles of the CGLCG motif in the D1- and D3-domains. J Biol Chem. 1998; 273 (51):34527–34. [PubMed: 9852122]
20. Wickstrom C, Carlstedt I. N-terminal cleavage of the salivary MUC5B mucin. Analogy with the Van Willebrand propolypeptide? J Biol Chem. 2001; 276 (50):47116–21. [PubMed: 11602588]
21. Huang RH, Wang Y, Roth R, Yu X, Purvis AR, Heuser JE, Egelman EH, Sadler JE. Assembly of Weibel-Palade body-like tubules from N-terminal domains of von Willebrand factor. Proc Natl Acad Sci U S A. 2008; 105 (2):482–7. [PubMed: 18182488]
22. Wickstrom C, Christersson C, Davies JR, Carlstedt I. Macromolecular organization of saliva: identification of ‘insoluble’ MUC5B assemblies and non-mucin proteins in the gel phase. Biochem J. 2000; 351(Pt 2):421–8. [PubMed: 11023828]
23. Lidell ME, Hansson GC. Cleavage in the GDPH sequence of the C-terminal cysteine-rich part of the human MUC5AC mucin. Biochem J. 2006; 399 (1):121–9. [PubMed: 16787389]
24. Escande F, Lemaitre L, Moniaux N, Batra SK, Aubert JP, Buisine MP. Genomic organization of MUC4 mucin gene. Towards the characterization of splice variants. Eur J Biochem. 2002; 269 (15):3637–44. [PubMed: 12153560]
25. Lin HH, Chang GW, Davies JQ, Stacey M, Harris J, Gordon S. Autocatalytic cleavage of the EMR2 receptor occurs at a conserved G protein-coupled receptor proteolytic site motif. J Biol Chem. 2004; 279 (30):31823–32. [PubMed: 15150276]
26. Kesimer M, Scull M, Brighton B, DeMaria G, Burns K, O’Neal W, Pickles RJ, Sheehan JK. Characterization of exosome-like vesicles released from human tracheobronchial ciliated epithelium: a possible role in innate defense. FASEB J. 2009; 23 (6):1858–68. [PubMed: 19190083]
27. Carson DD. The cytoplasmic tail of MUC1: a very busy place. Sci Signal. 2008; 1(27):pe35. [PubMed: 18612140]
28. Wen Y, Caffrey TC, Wheelock MJ, Johnson KR, Hollingsworth MA. Nuclear association of the cytoplasmic tail of MUC1 and beta-catenin. J Biol Chem. 2003; 278 (39):38029–39. [PubMed: 12832415]
29. Yin BW, Lloyd KO. Molecular cloning of the CA125 ovarian cancer antigen: identification as a new mucin, MUC16. J Biol Chem. 2001; 276 (29):27371–5. [PubMed: 11369781]
30. Khatri IA, Wang R, Forstner JF. SEA (sea-urchin sperm protein, enterokinase and agrin)-module cleavage, association of fragments and membrane targeting of rat intestinal mucin Muc3. Biochem J. 2003; 372 (Pt 1):263–70. [PubMed: 12605599]
31. Palmai-Pallag T, Khodabukus N, Kinarsky L, Leir SH, Sherman S, Hollingsworth MA, Harris A. The role of the SEA (sea urchin sperm protein, enterokinase and agrin) module in cleavage of membrane-tethered mucins. FEBS J. 2005; 272 (11):2901–11. [PubMed: 15943821]
32. Kesimer M, Makhov AM, Griffith JD, Verdugo P, Sheehan JK. Unpacking a gel-forming mucin: a view of MUC5B organization after granular release. Am J Physiol Lung Cell Mol Physiol. 2010; 298 (1):L15–22. [PubMed: 19783639]

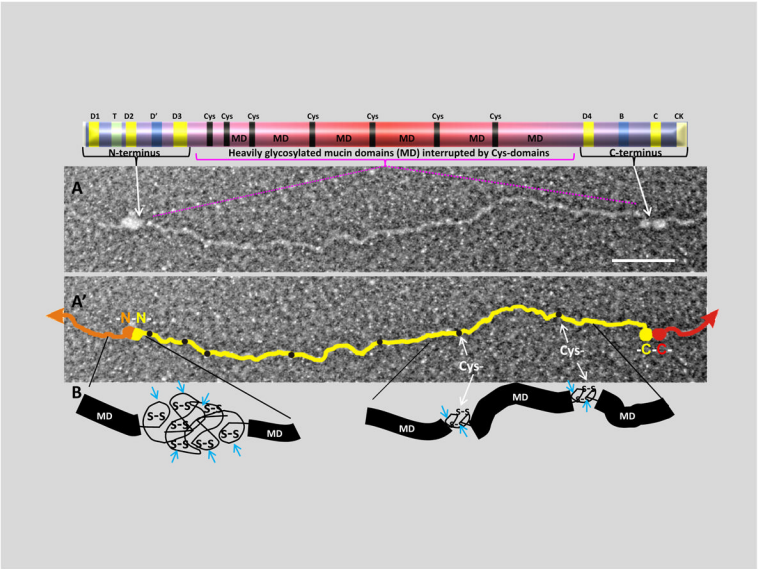


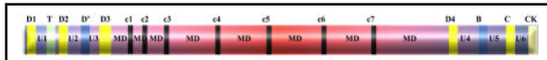
Figure 1. Illustration of the regions of MUC5B (AC) accessible/inaccessible to proteolysis and possible cleavages

A)- An electron microscopy image of a typical linear MUC5B (AC) molecule with a typical structural domain representation of the MUC5B subunit on the top panel. The mucin sample was isolated, prepared, and observed in an electron microscopy as described previously in detail [32]. The MUC5B (AC) molecule is assembled from multiple large subunits via disulfide bond (**S-S**) mediated dimerization between COOH-terminal (**C-**) domains of monomers to form dimers and subsequent oligomerization via NH_2-terminal (**N-**) domains to form higher oligomers. The mucin was traced with color-coded lines for individual subunit in **A′**. The assignment of the structure as a monomer (yellow trace) is made on the basis of the length (~400–600 nm). The **N-N** terminal region and **C-C** terminal region can be delineated by their size. Scale bar, 50 nm.

B)- Naked **N-** and **C-** terminal protein regions where the molecule makes oligomerisation, are susceptible to proteolysis (blue arrows). Although the cystein rich regions (Cys-) are located inside the highly glycosylated mucin domains, they are also targets for proteases. Despite these cleavage sites, mucin is still intact in normal conditions because they are held together by disulfide bonds

Table 1

Peptides identified from MUC5B.

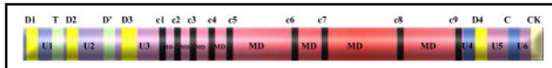


Peptide identified	Domain	Peptide identified	Domain
VCSTWGDFHYK, TFDGDVFR, FPGLCNYVFSEHCR, AAYEDFNVQLR RGLVGSRPVVTR, EELPYSR, TGLLVEQSGDYIK, LVLTFLWNGEDSALLELDPK, YANQTCGLCGDFNGLPAFNEFYAHNAR	**D1**	WSSWYNGHRPEPGLGGGDFETFENLR GYQVCPVLADIECR AAQLPDMPLEELGQQVDCDR GLMCANSQQSPPLCHDYELR	**C1**
LTPLQFGNLQK, TLLGPAFAECHALVDSTAYLAACAQDLCR CPTCPCATFVEYSR, QCAHAGGQPR, NWRCPELCPR	**U1**	CQWTEWFDEDYPK, SEQLGGDVESYDK AAGGHLCQQPK, DIECQAESFPNWTLAQVGQK VHCDVHFGLVCR, NWEQEGVFKMCYNYR, VLCCSDDHCR	**C2**
TCPLNMQHQECGSPCTDTCSNPQR AQLCEDHCVDGCFCPPGSTVLDDITHSGCLPLGQCPCTHGGR	**TIL**	CEWTEWFDVDFPTSGVASGDMETFENIR MCWAPK, AENYPEVSIDQVGQVLTCSLETGLTCK, NEDQTGR, FNMCFNYNVR	**C3**
LYDLHGDCSYVLSK, CADSSFTVLAELR, CGLTDNENCLK AVTLSLDGGDTAIR, LDPAHQGQMCGLCGNFNQNQADDFTALSGVVEATGAAFANTWK	**D2**	SYPMPGPSGGDFDTYSNIR AAGGAVCEQPLGLECR, AQAQPGVPLR ELGQVVECSLDFGLVCR, EQVGKFK, MCFNYEIR	**repeats at C4, C5, C6**
AQAACANAR, NSFEDPCSLSVENENYAR, HWCSRLTDPNSAFSR CHSIINPKPFHSNCMFDTCNCER, SEDCLCAALSSYVHACAAK, GVQLSDWR, DGVCTK, YMQNCPK YAYVVDACQPTCR, GLSEADVTCSVSFVPVDGCTCPAGTFLNDAGACVPAQECPCYAHGTVL- APGEVVHDEGAVCSCTGGK, LSCLGASLQK,	**U2**	PGVPLGELGQVVECSLDFGLVCR	**C6**
		AFGQFFSPGEVIYNK, AGCHFYAVCNQHCDIDR, VVLLDPK PVANVTCVNK, HLPIK, VSDPSQPCDFHYECECICSMWGGSHYSTFDGT-SYTFR, ALSIHYK, EEGLILFDQIPVSSGFSK, NGVLVSVLGTTTMR VDIPALGVTVTFNGQVFQAR LPYSLFHNNTEGQCGTCTNNQR, DDCLQR DGTTAASCK, DMAKTWLVPDSR	**D4**
VDCNTCTCR, WECSHR, LCLGTCVAYGDGHFITFDGDR	**D′**	GRLEVPCQSLEAYAELCR, GVCSDWR GATGGLCDLTCPPTK, VYKPCGPIQPATCNSR, NQSPQLEGMAEGCFCPEDQILFNAHMGICVQA- CPCVGPDGFPK	**U4**
IVTENIPCGTTGTTCSK, LFVESYELILQEGTFK, GPGGDPPYK, KTSVFIR, LHQDYKGR, VCGLCGNFDDNAINDFATR, SVVGDALEFGNSWK, LSPSCPDALAPK, DPCTANPFRK, QCSILHGPTFAACR	**D3**	PLPCDAQGQPPPCNRPGFVTVTRPR	**U5**
ACHDAGLCVSWR, TPDTCPLFCDFYNPHGGCEWHYQPCGAPCLK, NPSGHCLVDLPGLEGCYPK, CPPSQPFFNEDQMK, CVAQCGCYDK, DGNYYDVGAR, SLEACTCTYEDR	**U3**	KTGCCYSCEEDSCQVR, YSAEAQAMQHQCTCCQER, GFPAQEATAV	**CK**

D: Von Willebrand Factor D like domain, **U:** Unique region, **T:** trypsin inhibitory like domain, **c:**cysteine-rich region, **MD:** heavily *O*-glycosylated mucin domains, **B:** Von Willebrand Factor like B domain, **C:** Von Willebrand Factor like C domain, **CK:** cysteine-knot domain.

Table 2

Peptides identified from MUC5AC.



Peptide identified	Domain	Peptide identified	Domain
GPSGVPLR, VCSTWGSFHYK, TFDGDVFR, FPGLCNYVFSEHCGAAYEDFNIQLR, SQESAAPTLSRVLMK, VDGVVIQLTK, GSVLVNGHPVLLPFSQSGVLIQQSSSYTK TCGLCGDFNGMPVVSELLSHNTK	D1	DEGYTFCESPR, AESFPNTPLGR GQDVICSHTEGLICLNK, NQLPPICYNYEIR IQCCETVNVC	C4
LTPMEFGNLQK, QDLCFCEDTDLLSCVCHTLAEYSR QCTHAGGLPQDWR, GPDFCPQK	U1	CTWTTWFDVDFPSPGPHGGDKE, TYNNIIR RPEEITR, SHPEVSIEHLGQVVQCSR, EEGLVCR NQDQQGPFK, MCLNYEVR, VLCCETPK	C5
CPNNMQYHECR, SPCADTCSNQEHSR ACEDHCVAGCFCPEGTVLDDIGQTGCVPVSK	TIL	WFDVDFPSPGPHGGDK, ETYNNIIR RPEEITR, AESHPEVSIEHLGQVVQCSR EEGLVCR, NQDQQGPFK, MCLNYEVR, VLCCETPK	Repeats C6, C7, C8, C9
WSCQEVPCPDTCSVLGGAHFSTFDGK QYTVHGDCSYVLTKPCDSSAFTVLAELR CGLTDSETCLK, LRGQTCGLCGNFNSIQADDFR	D2	LYPAGSTIYRHR, DLAGHCYYALCSQDCQVVR	U4
TLSGVVEATAAAFFNTFK, TQAACPNIR NSFEDPCSLSVENEK, YAQHWCSQLTDADGPFGR SEDCLCAALSSYVHACAAK, GVQLGGWR DGVCTKPMTTCPK, SMTYHYHVSTCQPTCR, SLSEGDITCSVGFIPVDGCICPK, GTFLDDTGK, CVQASNCPCYHR, GSMIPNGESVHDSGAICTCTHGK LSCIGGQAPAPVCAAPMVFFDCR, SCHTLDMTCYSPQCVPGCVCPDGLVADGEGGCITA-EDCPCVHNEASYR	U2	VEKPTCANGYPAVK, VLVDNYFCGAEDGLSCPR, SIILEYHQDR, KPVHGVMTNEIIFNNK, FANNTEGQCGTCTNDR	D4
VGCNTCTCDSR, DSTQDSFR, VVTENVPCGTTGTTCSK, IFLGGFELK, TSIFINLSPEFK, VCGLCGNFDDIAVNDFATR	D3	VFEPCHTVIPPLLFYEGCVFDR, CHMTDLDVVCSSLELYAALCASHDICIDWR, TGHMCPFTCPADK,	U5
SVVGDVLEFGNSWK, LSPSCPDALAPK, DPCTANPFRK, QCSILHGPTFAACHAHVEPAR, TPSICPLFCDYYNPEGQCEWHYQPCGVPCLR, GDCLRDVR, GLEGCYPK, CPPEAPIFDEDK, MQCVATCPTPPLPPR, SYRPGAVVPSDK NCQSCLCTER, GVECTYK, AEACVCTYNGQR, FHPGDVIYHTTDGTGGCISAR, CGANGTIER	U3	CLGPHGEPVK, VGHTVGMDCQECTCEAATWTLTCRPK, LCPLPPACPLPGFVPVPAAPQAGQCCPQYSCACNTSR,	C
TRLPTASASLPPVCGE, CLWSPWMDVSRPGR GTDSGDFDTLENLR, VCESPR, AEDAPGVPLR, VQCSPDVGLTCR, NREQASGLCYNYQIR	C1	CPAPVGCPEGAR, SPAHLFYPGETWSDAGNHCVTHQCEK, HQDGLVVVTTK, ACPPLSCSLDEAR, SLIIQQQGCSSSEPVR, GNCGDSSSMYSLEGNTVEHR CQCCQELR	U6
DEGYTFCESPR, AESFPNTPLADLGQDVICSHTEGLICLNK, NQLPPICYNYEIR, IQCCETVNVCR,	C2	TSLRNVTLHCTDGSSR, AFSYTEVEECGCMGR, CPAPGDTQHSEEAEPEPSQEAESGSWER	CN
WFDVDFPSPGPHGGDK, ETYNNIIR RPEEITR, AKSHPEVSIEHLGQVVQCSR, EEGLVCR, NQDQQGPFK MCLNYEVR, VLCCETPR,	C3		

D: Von Willebrand Factor D like domain, **U:** Unique region, **T:** trypsin inhibitory like domain, **c:** cysteine-rich region, **MD:** heavily *O*-glycosylated mucin domains, **B:** Von Willebrand Factor like B domain, **C:** Von Willebrand Factor like C domain, **CK:** cysteine-knot domain.

Table 3

Peptides identified from membrane tethered mucins:

Peptides identified MUC1	**Domain**	**Peptides identified MUC16**	**Domain**
DISEMFLQIYK, QGGFLGLSNIK EGTINVHDVETQFNQYK	**SEA domain**	SPSQVSSSHPTR, EGTSLGLTPLNTR, SAQFSSSHLVSELR VTMSSTFSTQR, SLSMGNSTHTSMTDTEK LTTLFSTGQAARSGSSSSPISLSTEK, EPSISPEIR EPSISPEIRSTVR, SSGVTFSRPDPTSK, SSDSPSEAITR SSKTTR, EIFPSINTEETNVK, REPTYFLTPR, AFTAATTEVSR SVTMLSTFAGLTK, TIATQTGPHR, ESYSSVPAYSEPPKVTSPMVTSFNIR, IIKFPTSPILAESSEMTIK VSTGATTEVSR, EDVTSIPGPAQSTISPDISTR, GPEDVSWPSRPSVEK, LSTSPIK, NMTIPTLTLSPGEPK, ELQGLLKPLFR, LASLRPEK, NSLYVNGFTHR, NTSVGPLYSGCR, LTLLRPEK, DGAATGVDAICTHR, EQLYWELSK, LTNDIEELGPYTLDR, VLQGLLGPIFK, DGAATGVDAICIHHLDPK, LYWELSQLTNGIK, ELGPYTLDR YEEDMHRPGSR, VLQTLLGPMFK, SEKDGAATGVDAICTHR SPGLDR, EQLYWELSQLTNGIK, VLQGLLGPMFK, VLQGLLKPLFK, STSVGPLYSGCR, NGAATGMDAICSHR EQLYWELSQLTHGIK, DGAATGVDAICTHHLNPQSPGLDR EQLYWQLSQMTNGIK, VLQGLLSPIFK, LTSLRPEK DGAATGMDAVCLYHPNPK, RPGLDR, LTLLRPEKHEAATGVDTICTHR, HEAATGVDTICTHR VDPIGPGLDR, LYWELSQLTNSITELGPYTLDR, LTLLRPK, KDGAATGVDAICTHR, VLQGLLRPLFK HGAATGVDAICTLR, LDPTGPGLDR, LYWELSQLTNSVTELGPYTLDR, DSLYVNGFTHR VDAVCTHRPDPK, LSQLTHGITELGPYTLDR, YEENMHHPGSR, VLQGLLRPVFK, LTTLRPKK, VDAICTYRPDPK, LTLLRPEKR, RGAATGVDTICTHR, LDPLNPGLDR, GIIELGPYLLDR, GSLYVNGFTHR, YEEDMHCPGSR, VLQSLLGPMFK, YEENMQHPGSR, ESLYVNGFHPR, ISLYVNGFTHR, DSLYVNGFTQR, NTSIGPLYSSCR, VDAICTHHPDPQSPGLNR, VDAICTHRPDPK, VLQGLLMPLFK, VLQGLLR, DGTATGVDAICTHHPDPK, YEENMWPGSR, DGEATGVDAICTHRPDPTGPGLDR, YMADMGQPGSLK HLLSPLFQR, VDLLCTYLQPLSGPGLPIK, QVFHELSQQTHGITR, LGPYSLDK, SSMGPFYLGCQLISLRPEK, DSLFINGYAPQNLSIR VTTLYK, GSQLHDTFR, ALFSSNLDPSLVEQVFLDK, AQPGTTNYQR, NIEDALNQLFR, SYFSDCQVSTFR, HHTGVDSLCNFSPLAR, SSVLVDGYSPNR, EGEYNVQQQCPGYYQSHLDLEDLQ, GAATGVDTICTHR DSSATAVDAICTHRPDPEDLGLDR, TLLRPEKR, LTLLRPEKDK SPGLNR, GEATGVDAICTHR, LSQLTHSITELGPYTLDR LYWELSNLTNGIQELGPYTLDR, NTSVGLLYSGCR, VAIYEEFLR	**SEA domain**
NYGQLDIFPAR, DTYHPMSEYPTYHTHGR YVPPSSTDRSPYEK, VSAGNGGSSLSYTNPAVAAASANL	**Cytoplasmic tail**		
MUC4			
GVSLFPYGADAGDLEFVR, TVDFTSPLFK PATGFPLGSSLR, DPVALVAPFWDDADFSTGR GTTFYQEYETFYGEHSLLVQQAESWIR, SGNPVLMGFSSGDGFFENSPLMSQPVWER, FLNSNSGLQGLQFR, LHREERPNYR, EERPNYR LECLQWLK	**C-terminal-2 (CT2) domain**		
FQPVSIGR, WGLGSRQLCSFTSWR, GGVCCSYGPWGEFR, EGWHVQR, PWQLAQELEPQSWCCR, WNDKPYLCALYQQR	**CT3-Cystein rich domain**		
TAQTGSAQATNFIAFAAQYR, SSSLGPVTVQWLLEPHDAIR, TEGLLGVWNNNPEDDFR	**CT4/vWFD domain**		
SLEPFTLEILAR, IGLASALQPR	**CT5 (N-glycosylation rich domain)**		
DGGTFGR, YCEGSEDACEEPCFPSVHCVPGK LGTLDMR, IDSAAPASGSPIQHWMVISEFQYRPR	**CT6 (Cys rich domain)**		
FLAGNNFSPTVNLELPLR, NDVVFQPISGEDVR, DVTALNVSTLK	**CT8 domain**		
GYDLVYSPQSGFTCVSPCSR, GYCDHGGQCQHLPSGPR, FSYFLNSAEAL	**EGF2 domain**		
MUC20			
LSVASPEDLTDPR	**C- terminal**		

MD: mucin domain, **CT:** C-terminal domains, **T:** transmembrane, **Cy:** cytoplasmic domain, **SEA:** (**S**ea urchin sperm protein, **E**nterokinase and **A**grin) domain.

Table 4

List of non-specific cleavage peptides detected by mass spectrometry after MASCOT “error tolerant” database search. The fragments are colored in red. The MS/MS spectra of all of these peptides are supplied at supplementary data 2.

Mucin	Cleavage site	Domain	Comment	Digest #
MUC5B	FP/GLCNYVFSEHCR;	vWFD1	non-specific cleavage	1,2,3
	YA**NQT**CGL/CGDFNGLPAFNEFYAHNAR;	vWFD1	Putative N-glycosylation at the downstream non-specific cleavage cuts a putative multimerization site	3
	YA**NQT**CGLCGDFNGL/PAFNEFYAHNAR;	vWFD1	Putative N-glycosylation at the downstream	3
	FAECH/ALVDSTAYLAACAQDLCR;	N-terminal unique	non-specific cleavage	1,2,3
	ALVDST/AYLAACAQDLCR;	N-terminal unique	non-specific cleavage	1,2,3,
	ALVDST AY/LAACAQDLCR;	N-terminal unique	non-specific cleavage	1,
	T/CPLNMQHQECGSPCTDTCSNPQR;	N-Terminal TIL	non-specific cleavage	1,2,3,
	QADDFTAL/SGVVEATGAAFANTWK;	vWFD2	Chymotrypsin type cleavage	1,2
	NSFED/PCSLSVENENYAR;	N-terminal unique	non-specific cleavage	1,2,3
	SFV/PVDGCTCPAG;	N-terminal unique	non-specific cleavage	2
	CHSIINPK/PFHSNCMFDTCNCER;	N-terminal unique	LysC type cleavage	1,2,3
	VCGL/CGNFDDNAINDFATR	vWFD3	Chymotrypsin type cleavage	1,2,3
	SV/VGDALEFGNSWK;	N-termimal MUC11p15	non-specific cleavage	1
	FCTAVAAYAQ/ACHDAGLCVSWR	N-terminal unique	non-specific cleavage	1,2,3
	NPSGH/CLVDLPGLEGCYPK;	N-terminal unique	non-specific cleavage	
	ENCQSCNCTPSGIQCAH/SLEACTCTYEDR;	N-terminal unique	non-specific cleavage	1,2,3
	MPLEE/LGQQVDCDR;	Cys-rich-1	non-specific cleavage	2,3
	VL/CCSDDHCR;	Cys-rich-2	Chymotrypsin type cleavage	1,3
	ELGQVVEC/SLDFGLVCR;	Cys-rich-4,5,6,7.	non-specific cleavage	1,2,3
	SEWLDY/SYPMPGPSGGDFDTYSNIR;	Cys-rich-4,5,6,7.	Chymotrypsin type cleavage	1,2,3
	AGCHFY/AVCNQHCDIDR	C-terminal MUC11p15	non-specific cleavage	1,3
	AQAQ/PGVPLGELGQVVECSLDFGLVCR;	Cyst-rich-7	non-specific cleavage	1,2
	ECECICSMWGGSHY/STFDGTSYTFR	vWFD4	Chymotrypsin type cleavage	1,2,3
	SCVCDEGSVSVQCK/PLPCDAQGQPPPCNR/PGFVTVTRPR;	vWFD4	non-specific cleavage	1
	NLSLYLD/NHYCTASATAAAAR;	vWFD4	Putative N-glycosylation at the downstream Glu- C type cleavage	3
	DQILFNAH/MGICVQACPCVGPDGFPK;	C-terminal unique	non-specific cleavage	1,2,3
MUC5AC	SP/CADTCSNQEHSR;	N-Terminal TIL	NTCB type cleavage	1,2,3
	GSMI/PNGESVHDSGAICTCTHGK;	N-terminal unique	non-specific cleavage	1,2,3
	SE/DCLCAALSSYVHACAAK;	N terminal unique	available also in MUC5B, Glu- C type cleavage	1
	PSL/RTIPVVR;	N terminal unique	Chymotrypsin type cleavage	2,3
	WKLS/PSCPDALAPK;	N terminal unique	non-specific cleavage	1,2,3

Mucin	Cleavage site	Domain	Comment	Digest #
	KQCSIL/HGPTFAACHAHVEPAR;	N terminal unique	Chymotrypsin type cleavage	1,2,3
	SYRPGAVV/PSDKNCQSCLCTER;	N terminal unique	non-specific cleavage	1,2,3
	PGAVV/PVDGCICPK;	N terminal unique	non-specific cleavage	2,3
	GLICLNKNQL/PPICYNYEIR;	Cys-rich-2,4	Chymotrypsin type cleavage	1
	KWFDV/DFPSPGPHGGDK;	Cys-rich-3,5,6,7,8,9	non-specific cleavage	1,2,3
	SI/ILEYHQDR;	vWFD4	non-specific cleavage	1,2
	LY/PAGSTIYR;	C-terminal unique	Chymotrypsin type cleavage	1,,3
	FCPEGMTLF/STSAQVCVPTGCPR;	C terminal unique	Chymotrypsin type cleavage	1,2,3
	AC/PPLSCSLDEAR;	C-terminal unique	non-specific cleavage	1,2,3
	TVG/MDCQECTCEAATWTLTCRPK;	C-terminal unique	non-specific cleavage	2
	TC/PRVEKPTCANGYPAVK;	C-terminal unique	non-specific cleavage	1,3
MUC4	SSSL/GPVTVQWLLEPHDAIR; SSSLG/PVTVQWLLEPHDAIR;	vWFD (CT4)	Chymotrypsin type cleavage	1,2
	SL/EPFTLEILAR; SLE/PFTLEILAR;	N-Glycosylation rich (CT5)	Chymotrypsin type cleavage	1,2
MUC1	MSEY/PTYHTHGR;	Intracellular	non-specific cleavage	1,2,3
	YV/PPSSTDR;	Intracellular	non-specific cleavage	2,3
	DTYHPMS/EYPTYHTHGR;	Intracellular	non-specific cleavage	1,2
MUC16	SEKDGA/ATGVDAICTHR;	SEA	non-specific cleavage	1,2,3
	LYWE/LSQLTH GIKELGPYTLDR;	SEA	Glu-C type cleavage	1,2